Team Bloom
Fraud Detection with Medicare Provider Data

Data Cleaning
Data Formatting
Data Exploration

Austin Knies (Economics)

Machine Learning Operations

Jonathan Leslie (Economics)

Problem: Health care fraud creates **billions of dollars in costs** every year.

<u>What is health care fraud?</u>

When characteristics of medical provision are **misrepresented** or **misstated** such that **unauthorized payments** are made

(DOJ)

<u>Who can create fraud?</u>

Individuals

- Doctors
- Patients

Institutions

- Health care organizations
- Health systems

(FBI, CMS)

<u>Examples of health care fraud:</u>

False or misrepresented claims

- Double billing
- Phantom billing
- Upcoding
- Unbundling

Identity theft

- Forgery
- Impersonation

(FBI, CMS)

<u>Who does fraud impact?</u>

- Patients
- Care systems
- Organizations
- Insurance companies
- Government

(FBI, CMS)

Sources: *FBI*, *CMS*, *DOJ*

Reducing health care fraud will save billions of dollars annually for insurers.

Identifying health care fraud is difficult and costly to do.

What we provide:

1. A model for **identifying potential fraud** in a large, unlabeled dataset of medical providers
   a. *Unsupervised learning*
2. A model for **classifying whether a particular provider** is potentially fraudulent or not
   a. *Supervised learning*

Data Overview

Medicare Physician & Other Practitioners by Provider and Service Datasets
Centers for Medicare and Medicaid Services (CMS) Public Use Files (PUFs)
2013 through 2020

Data.CMS.gov
Centers for Medicare & Medicaid Services

Information on health care utilization, payments, and charges by provider, service, and place of service

Derived from CMS administrative claims data for Original Medicare Part B beneficiaries

https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider-and-service

Data Cleaning and Formatting

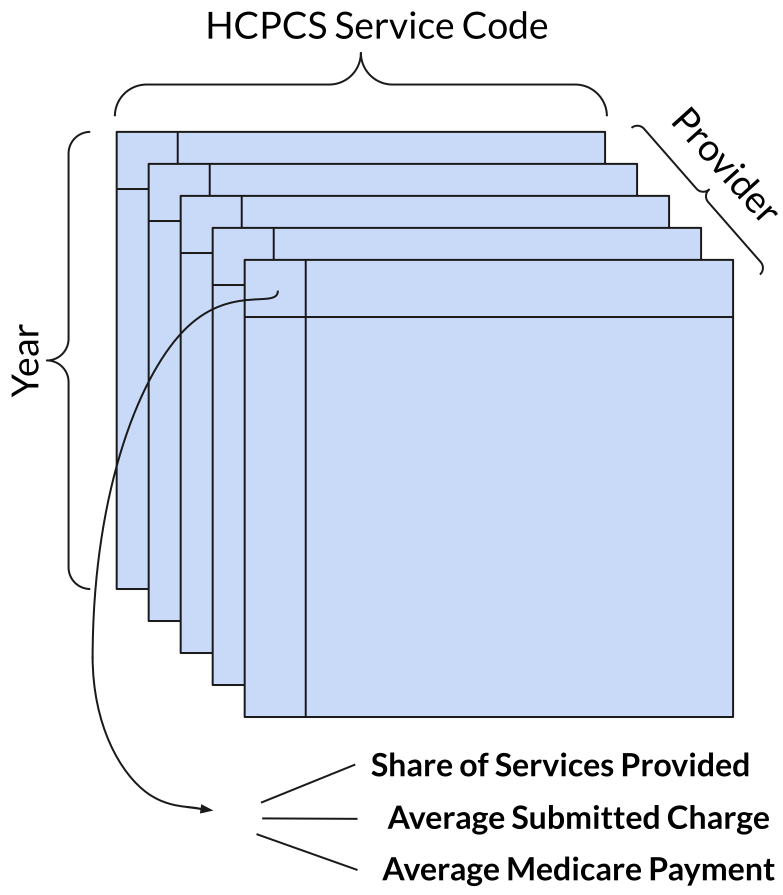
Over 77 million provider-year-HCPCS-POS combinations
1.5 million unique providers from 2013 to 2020

Subset of data we look at:

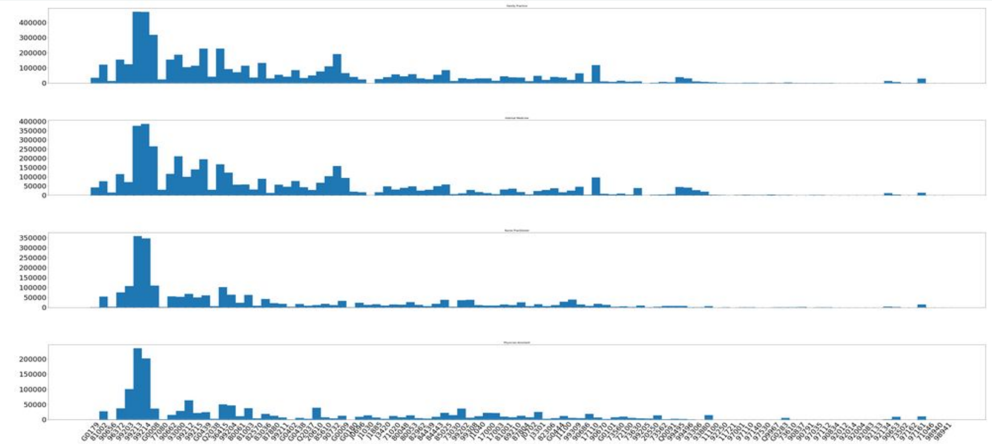
Non-facility (office) place of service
Balanced, type-constant panel of providers
Top 100 HCPCS service codes by provider-year frequency
Family Practice, Internal Medicine, NP, and PA providers

Artificial data set with full shape:

100,645 unique providers
8 years of data
<u>100 HCPCS service codes</u>
80,516,000 total observations



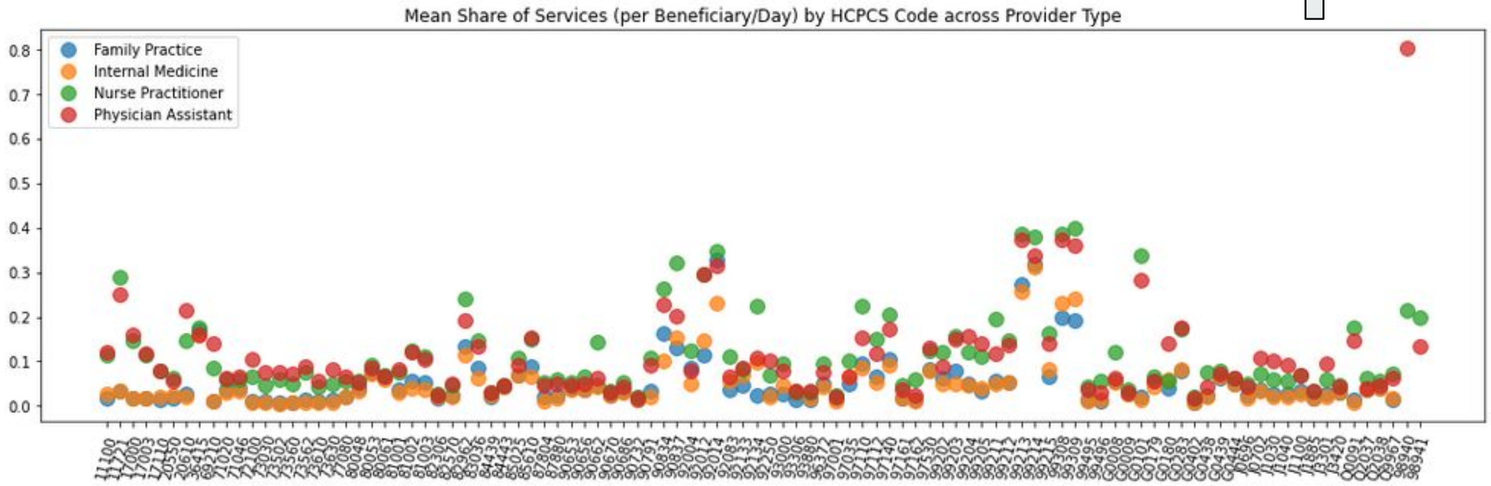
Data Exploration



Similarities across provider types in distribution of service code frequency

Accounting for heterogeneity will enable improved detection of anomalous/fraudulent activity

Provider type heterogeneity contributes to **differences** in share of total services and charges submitted and paid



Machine Learning Operations

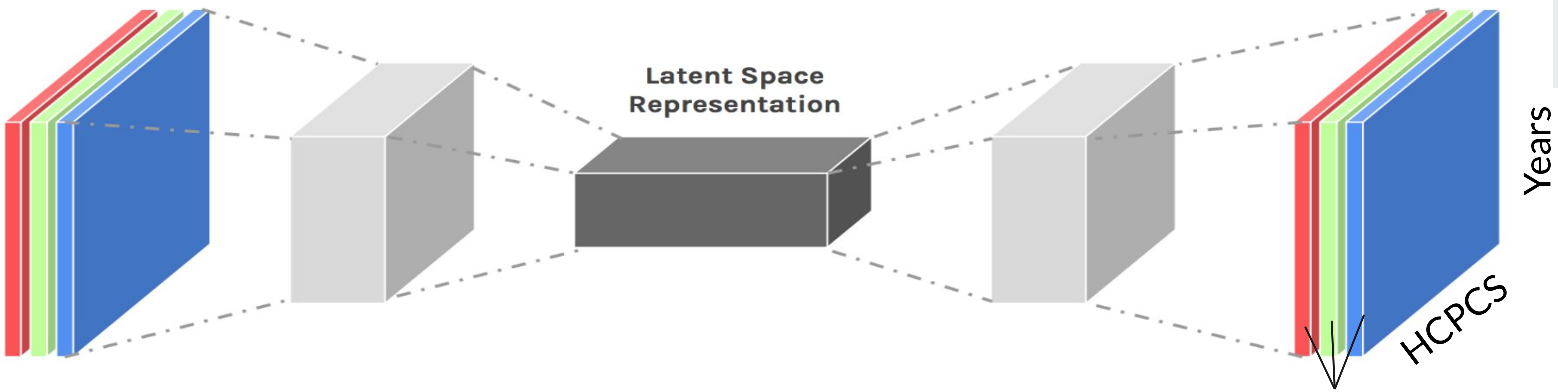
Goal: <u>Create a binary classifier that takes inputs about a medical provider and outputs whether or not the provider is potentially fraudulent</u>

Two Main Issues:

1. The data is *unlabeled* - need to first designate what is possible fraud and what isn't
2. Fraud is expected to be *very uncommon* - issue of *unbalanced classes* when predicting fraud

Three Main Processes:

1. Classify *what constitutes fraud* using an **Autoencoder** model
2. Generate *fake fraudulent examples* to balance the classes using a **Generative Adversarial Network**
3. Train a *binary classifier* to predict fraud given inputs using a **Convolutional Neural Network** model



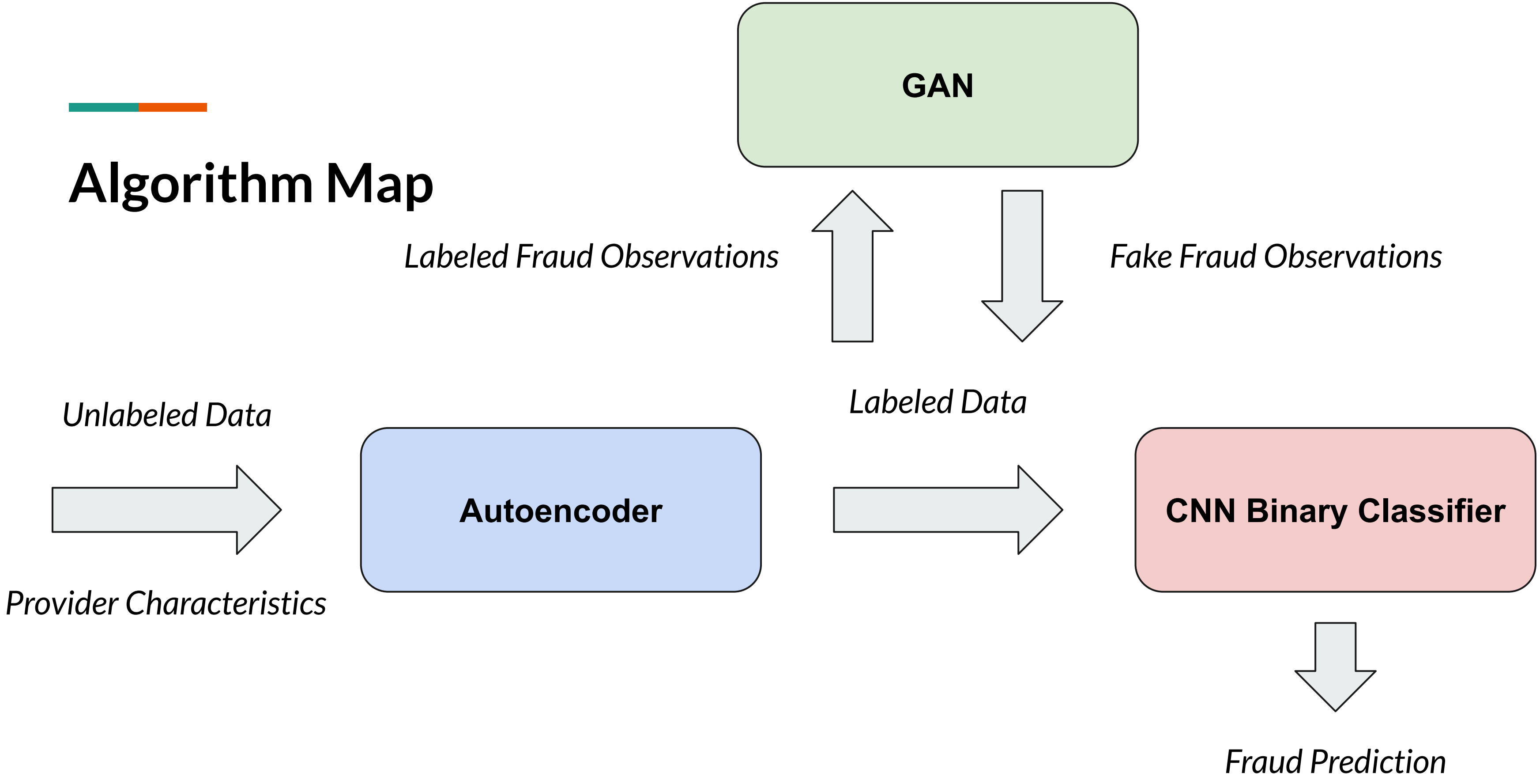
Autoencoder

- For each provider, observations are:
- **(years, HCPCS, variables of interest)**
- Learns to *recreate the input* after shrinking it to a smaller *latent space*
- 3D observations, so *convolutional neural network layers*
- Label observations as fraudulent with the highest *reconstruction errors*
- Anomalous if autoencoder can't recreate well based on learned data distribution

GAN

- Two *convolutional neural networks*
- One generates convincing *fake data*
- Other distinguishes *real* from *fake* data
- *Generator* gets better at creating fake data that *resembles the real data*
- Create *fake fraud observations* that look like the *real fraud* (as labeled by the Autoencoder)
- Add *fake fraud* to original data to improve classification - *balanced classes*

Algorithm Map



CNN Classifier Results

Data

- Limit to *Internal Medicine* and *Family Physician* provider types
- Limit to top 12 HCPCS service codes

Autoencoder

- Label observations with highest 0.5% *reconstruction errors* as fraudulent

GAN

- Generates 20,000 fake fraud observations (~75,000 original observations)

Actual \ Predicted	Not Fraud	Fraud
Not Fraud	*Validation Data* w/ Fake: 37,024 w/o Fake: 37,039	*Validation Data* w/ Fake: 22 w/o Fake: 7
Fraud	*Validation Data* w/ Fake: 81 w/o Fake: 121	*Validation Data* w/ Fake: 104 w/o Fake: 64

Conclusion

- Medical fraud results in billions of dollars in extra costs for payers, including federal and state governments
- We produce a general modeling framework for anomaly detection and prediction of rare class types (e.g. fraud) with unlabeled data
- Methodology can be used to try to help detect and prevent future medical fraud in the Medicare program
- Future Improvements:
    - Hyperparameter optimization to improve prediction accuracy
    - Compare predicted fraud by provider type, location, etc.
    - Application to other data sources

Links and Resources

LinkedIn Pages:

https://www.linkedin.com/in/austinknies/

https://www.linkedin.com/in/jonathan-leslie-2b4397201/

Annotated GitHub Repository

https://github.com/austinknies/fall22-bloom